

# A Guide to Time Series Databases - Dataversity

*Keith D. Foote*

Time series databases (or TSDBs) are databases that have been optimized for processing time series data. Time series data is made up of data records that are indexed using timestamps. The timestamps provide a reference for each of the data records and show how they relate to one another in time. An example of time series data would be the data taken from an assembly line sensor, which supplies a continuous flow of data, including when the data was recorded.

A time series database allows large amounts of time-stamped data to be stored in a format that supports complex analysis, quickly and efficiently.

Time series databases are often used to handle both financial data and tracking market fluctuations through the course of the day. Having the capability to match data points (identifiable elements, normally a numeric ID) with timestamps, on a massive scale, has allowed some savvy investors to predict trends and/or market anomalies, resulting in significant profits. ([Time series analysis](#) can also help in understanding the underlying reasons for trends.)

The primary benefit of time series databases is that they can be used to quickly analyze and identify patterns in the data.

## How Time Series Databases Work

Time series databases are scalable and capture a series of fixed values (the time) and a series of dynamic values (the changes that take place). For example, it could be considered acceptable when a piece of manufacturing equipment makes a mistake, on average, once every four hours. If the rate of mistakes increases to once an hour, that would be worth an investigation, and having the time listed when each mistake took place allows for an examination of what happened at that time.

Time series databases use sequences of data points containing two pieces of identification: a numeric value for ID purposes and a timestamp.

Because all time series data records are timestamped, the order of the data can be used to deliver it into a stream processing engine, which treats it as though it were a data stream. The primary goal of time series databases is to be fast, and using a fast stream processing engine is an excellent way to access the best current speeds.

## The Uses of Time Series Databases

As use of smart devices and the [Internet of Things](#) continues to increase, massive amounts of real-time data traffic are being generated, with literally millions of events and transactions being recorded each day. Using time series data allows people to make reasonably good predictions about the future.

Time series analysis can be very useful for analyzing yearly, seasonal, and monthly trends in sales.

Time series data is used in:

- **Pattern recognition:** There are several methods for using [pattern recognition](#) for time series databases. They typically first transform the data into a more common format. A machine learning algorithm is then used to find and classify the pattern. When visual pattern recognition is the goal, the data is first transformed into a picture.
- **Statistics:** In this situation, data points are recorded and stored at regular intervals during an established period of time, rather than intermittently. Time series analysis uses [statistical methods](#) to analyze the data and find patterns within it.
- **Econometrics:** Time series data can be combined with [econometrics](#), using statistical and mathematical models to predict future economic trends.
- **Control engineering:** An engineering discipline focused on control systems. When [control engineering](#) is used with time series data, it can predict behavior in controlled environments.
- **Signal processing:** A field of engineering that attempts to analyze digital and analog signals, in terms of time. A time series analysis is a form of [signal processing](#), with recorded data points at regular intervals.
- **Finances:** Some financial analysts are using time series data (stock price movements, a business' sales over time) to predict a [company's future performance](#).

## Time Series Data Concerns

There are a few concerns associated with time series data that users should be aware of. Ignoring these concerns will increase the probability of poor, inaccurate predictions. These same concerns apply to statistics in general. They are:

- **The Quantity Consideration:** With too few samples, the accuracy of predictions on complex issues can suffer significantly. To produce accurate forecasts, a reasonable number of samples are needed to capture the essential elements for a time series analysis. (Asking 10 people who they will vote for, and monitoring changes in their opinions, will not be enough to predict the results of a statewide election.)
- **The Aggregation Consideration:** *Aggregation* means a total, made up of different parts, or a “whole” made by combining several elements. A high level of aggregation (a wide variety of samples) typically results in more accurate predictions. Having many similar samples may not represent reality. (Using the time series data of Volkswagen drivers to represent all car drivers would produce inaccurate predictions.)
- **The Update Consideration:** This deals with situations that may require the continuous regular updating of forecasts to capture new information. (Think of weather forecasts.) If the update frequency is too sporadic, useful information can be missed.
- **The Horizon Consideration:** In this case, the horizon represents the future. The further ahead predictions are made, the more uncertain the predictions become. For a prediction to be accurate, the data must be relevant and dependable for an extended period of time. (Predicting how many people will be consistently riding bicycles in 30 years would be impossible because of the huge number of unknown factors.)

## A Mini History of Time Series Databases

Time series data is a statistical tool that uses recurring patterns to predict future events. Predicting the sun will rise tomorrow morning is fairly easy, but what time will it rise? The Egyptians were using sundials and “shadow clocks” before 1500 BC and were able to answer that question because they kept records. People have been using the basics of time series analysis for a very long time.

Aside from astronomical predictions, the first recorded, published effort to use time series data took place in 1662, when John Graunt, a 17th-century haberdasher in London, [published a book](#) titled “Natural and Political Observations... Made Upon the Bills of Mortality.” Graunt conducted a study of death records and was able to predict the probability of a person of a certain age dying before their next birthday.

Time series analysis has evolved significantly since Graunt published his study. Time series databases grew out of the desire to process financial data and track market fluctuations throughout the day. The first successful computer tool for working with time series data was the [round-robin database tool](#), developed in 1999. The first open-source database was OpenTSDB, presented in 2011. The highly popular open-source database InfluxDB was made available to the general public in 2013.

## Some Popular Time Series Databases

**InfluxDB:** This is a very popular open-source database. It can be used both in-house and in the cloud. It offers templates for a variety of useful templates. As an open-source database, [InfluxDB](#) has been a key factor in the growing use of time series databases.

**QuestDB:** Also open source, this is an SQL database. It uses a column structure for storing data and appends new data to the bottoms of each column, recording the time order of the incoming data. [QuestDB](#) can also support relational modeling with time series data (you can write joins, while using SQL queries to read the data).

**TimescaleDB:** This is also an open-source SQL database. It is essentially an extension that works with PostgreSQL. [TimescaleDB](#) can be downloaded and used in-house but can also be used in a variety of clouds through the use of a multi-cloud management platform, called Aiven.

**“Kdb+”:** Considered unique. [Kdb+](#) can be described as a columnar time series database supporting in-memory computing and relational modeling. It has been used by the high-tech trading industry for several years and is written in a programming language called k (making it unique). The k language is known for array-processing.

**Druid:** A time series database, but it can also be used for extremely fast aggregations of time-ordered data. It can be described as a time-based analytics database. [Druid](#) comes with time-based partitions and compressed bitmap indexes for pruning data that is not needed. It uses a query language that is JSON-based. Druid also provides Druid SQL.

*Image used under license from Shutterstock.com*